

StarCoder: may the source be with you!

Year: 2023 | Citations: 1005 | Authors: Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov

Abstract

The BigCode community, an open-scientific collaboration working on the responsible development of Large Language Models for Code (Code LLMs), introduces StarCoder and StarCoderBase: 15.5B parameter models with 8K context length, infilling capabilities and fast large-batch inference enabled by multi-query attention. StarCoderBase is trained on 1 trillion tokens sourced from The Stack, a large collection of permissively licensed GitHub repositories with inspection tools and an opt-out process. We fine-tuned StarCoderBase on 35B Python tokens, resulting in the creation of StarCoder. We perform the most comprehensive evaluation of Code LLMs to date and show that StarCoderBase outperforms every open Code LLM that supports multiple programming languages and matches or outperforms the OpenAI code-cushman-001 model. Furthermore, StarCoder outperforms every model that is fine-tuned on Python, can be prompted to achieve 40% pass@1 on HumanEval, and still retains its performance on other programming languages. We take several important steps towards a safe open-access model release, including an improved PII redaction pipeline and a novel attribution tracing tool, and make the StarCoder models publicly available under a more commercially viable version of the Open Responsible AI Model license.